SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

DTIC FILE COPY

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
AD-A213 552 (S)	5 MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-1270
6a. NAME OF PERFORMING ORGANIZATION Natl. Inst. of Health	6b OFFICE SYMBOL (If applicable)
	7a NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research/NL
6c. ADDRESS (City, State, and ZIP Code) Bldg. 9, Room 1N107 Bethesda, MD 20892	7b ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING / SPONSORING ORGANIZATION AFOSR	8b OFFICE SYMBOL (If applicable) NL
	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-ISSA-88-0073
8c. ADDRESS (City, State, and ZIP Code) Building 410 Bolling AFB, DC 20332-6448	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2313; TASK NO. A5; WORK UNIT ACCESSION NO.

11 TITLE (Include Security Classification)

Unbiased Measures of Neuronal Information Transmission and Channel Capacity

12 PERSONAL AUTHOR(S)

Lance M. Optican, Timothy J. Gawne, Barry J. Richmond, Pinchas J. Joseph

13a. TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM 7/88 TO 8/88	16 June 89	40

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The response activity of nerve cells in the mammalian visual system was analysed for information bearing properties. Primary findings include: support for the multiplex filter hypothesis (nerve fibers encode multidimensional inputs by modulating the amplitudes of a few linearly independent temporal patterns summed to produce an output); the nature of encoding may permit recovery of an image's pattern regardless of its intensity or duration; in higher brain centers, encoded information becomes more evenly distributed among the temporal patterns of a nerve cell; higher brain centers, encoded information becomes more evenly distributed among the temporal patterns of a nerve cell; higher brain centers appear to affect the responses of lower centers through feedback; with sequentially changing visual inputs, independent responses can occur with images separated by about 30 msec.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. John F. Tangney	22b TELEPHONE (Include Area Code) (202) 767-5021 / 22c OFFICE SYMBOL NL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

FINAL REPORT

Perception and recognition of complex visual pictures depends on the normal function of a sequentially connected system of brain regions extending from the retina through inferior temporal cortex. The properties of these regions are derived from the function of the single neurons within them. Thus, to understand how visual perception occurs, we must learn how information is encoded by the neurons in these successive stages of processing. One clear consequence of such understanding would be the ability to predict the responses of single neurons to different pictures and to decode neuronal messages. We have been recording single neurons in several regions of the visual pathway with the goal of developing quantitative models of neuronal function. Such models should simulate the activity of single visual system neurons in response to visual stimuli. Over the past several years we have made substantial progress toward this goal. Our fundamental finding is that individual neurons in all the visual areas studied thus far (retinal ganglion cell fibers, lateral geniculate nucleus neurons, pulvinar neurons, primary visual cortical neurons, and inferior temporal cortical neurons) encode and transmit information about stationary, two-dimensional pictures that vary in form, brightness, and duration. The neurons use a multidimensional temporal code to represent and transmit their stimulus-dependent messages. Based on these findings we have begun to explore how the cooperation of these individual building blocks might give rise to higher visual cognitive functions such as perception, attention, and memory.

To study the messages carried by single visual system neurons more quantitatively than has been done previously, we developed a new approach to studying single neurons in which neurons are treated as communication channels that transmit information about pictures in their responses. To analyze a communication channel, a known set of signals is used as inputs, here visual stimuli, and then the responses are analyzed to find the representation of the input signal. To fully characterize the channel, the input set should cover the spectrum of all possible signals that can be encountered. In our application, this condition was satisfied by building a set of visual stimuli based on a complete orthogonal two-dimensional basis set of mathematical functions, the Walsh-Hadamard set. Each member of this set appears as a two-dimensional combination of black and white squares and rectangles within the stimulus border. This basic stimulus set can be considered an alphabet for pictures.

We adapted a well-known statistical technique, principal component analysis, to identify the optimal set of temporal patterns, the principal components, that describe the neuron's responses. Just as the Walsh patterns can be considered an alphabet for pictorial stimuli, the principal components can be considered an alphabet for the responses. To quantify the stimulus-response relation, we adapted and applied Shannon's information theory to the stimulus-response set. Our two-dimensional set of black and white Walsh patterns was defined as an input code, and the first three principal components were used to represent the responses as a temporally modulated output code. Our original analyses showed that neurons in both inferior temporal and primary visual cortex vary the strength and pattern of their responses independently.

Our original calculations used Shannon's basic formulation of information theory. We were able to show that stimulus-related information is carried in three or more simultaneous and independent patterns of activity or messages that are multiplexed onto the spike train. We learned that a response code based on temporal modulation carries at least twice as much information as a code based on response strength alone. However, because estimation of the stimulus-response probabilities from data which are continuous, noisy, or few in number (compared to the number of stimulus classes), such as are encountered in neurophysiological experiments, leads to biased overestimates of the transmitted information, we could not rely on our estimates of the absolute amount of information transmitted. To account for this problem over the past year we have used our newly acquired super-graphics workstation to develop an improved estimator of the amount of information transmitted under our conditions. Although computationally intensive, this new method allows a substantially more accurate assessment of the stimulus-related information transmitted in the responses of single neurons.

The improved estimator, T*, is corrected from the initial information estimate, T, by subtracting a bias term that is determined by calculating the information in the stimulus-response pairings both in the normal manner, T, and after the stimulus-response pairings have been randomized, B. If the stimulus-response relation in the experimental data is initially random, then the information calculated before and after the randomization will be equal; otherwise, the randomization procedure will yield a smaller calculated transmitted information than will the experimental data. Thus, the ratio of the information calculated after randomization to the information calculated before randomization, (B/T), is a factor that is sensitive to the data's signal-to-noise ratio, and can be used to correct for the bias: $T^* = T - (T/B)B$.

The new estimator, T*, is accurate to within 5% for sample sizes as small as 7 in simulated data sets contaminated with both Gaussian and uniform noise. For neuronal data, T*, even when calculated with as few as 7 samples per class, gave an accurate estimate of T calculated with 30 samples per class. Thus, we feel that T* is a better estimator of the actual transmitted information in biological signals since it minimizes error caused by quantization, noise, and small sample sizes.

Because our experiments are the first to systematically explore the whole extent of a visual pathway with similar methods, we are able to compare the ratio of temporal modulation information to response strength information over the whole extent of the occipitotemporal visual pathway. Originally we estimated that neurons in all the visual areas we studied transmitted about 1.0 bit of information in the temporal code, whereas it was only one-half to two-thirds that using a code based on response strength alone. Our improved information estimator, T*, shows that the information transmitted by single neurons in all the areas we have studied is between 0.4 and 0.6 bits; these lower values show the influence of the bias correction. Thus, the basic conclusion remains unchanged, i.e. information is carried in several independent temporally modulated patterns multiplexed into neuronal responses. In retinal ganglion fibers the ratio of information in the temporal code is 1.2 times as much as in the spike count, rises to 1.4 times in the lateral geniculate nucleus, 1.9 times in the pulvinar, 2.1 in striate cortex, and 2.4 in the inferior temporal cortex. Thus, temporal modulation plays an increasingly important role in information transmission at visual system stations located progressively further away from the retinal input.

Based on our informational analysis, we proposed multiplex-filter hypothesis of neuronal function. The hypothesis states that each neuron can be viewed as a small number (2-5) of simultaneously active spatial-to-temporal filters whose outputs are added or multiplexed together to form the response. Simple models based on this multiplex-filter hypothesis have predicted the temporally modulated responses of ganglion cell fibers, lateral geniculate neurons, and complex cells in striate cortex to arbitrarily constructed black and white patterns. Neither static receptive field models nor unidimensional response strength measures can correctly predict the temporally modulated responses of a neuron.

In the most common view of neuronal function, the strength of a neuron's response represents how closely the stimulus matches the receptive field's characteristics, e.g., orientation or color. Thus, if response strength were the only parameter a neuron could use to encode information, different stimulus features would be confounded by individual neurons. However, informational analysis showed previously that information about different stimulus parameters is not confounded but is carried separately across the different parts of the multidimensional neuronal code. Furthermore, although information theory does not require that the neural code be interpretable in terms of stimulus features, we have been able to show that the neural code can in fact be so interpreted. When the responses were represented by the first three principal components and plotted in a space whose axes are these three principal components, the responses elicited by an individual Walsh pattern appeared to lie near a single plane irrespective of duration or luminance.

In any one neuron's principal component space, many of the planes representing the many Walsh patterns appeared easily differentiable. This geometrical structure demonstrates that the generation of neuronal responses obeys certain rules, which form an intrinsic temporal neural code for visual features. A response could be decoded to determine the stimulus pattern irrespective of duration or luminance, if the plane into which the response falls could be ascertained. Information about duration and luminance would then be encoded relative to that plane. Since three points determine a plane, such a decoding scheme may require as few as three complementary neurons sharing related codes.

Now that we have found that temporal modulation of neuronal responses carries a substantial proportion of the stimulus-related information throughout the cortical visual system and have identified a potential set of rules that describe the neural code, we have begun to study how these neuronal codes might be generated in the lateral geniculate nucleus. In these experiments we are stimulating lateral geniculate neurons with the Walsh pattern set and compared their responses before and after reversibly inactivating the primary visual cortex by cooling. Cooling has caused two changes in lateral geniculate neuronal responses. First, it reduced the total information encoded in the responses; the proportion of the responses that represented temporal modulation was decreased significantly more than the response strength ($p < .05$, t-test). Second, cooling caused marked changes in the shapes and amplitudes of the response waveforms to some stimulus patterns, but not to others. These results suggest that feedback plays a central role in generating the normal neural codes for pictures throughout the visual system. Using our recently acquired computational graphics facilities we are now studying the effect

of these changes on the response representation revealed by the planes in the three-dimensional response graphs.

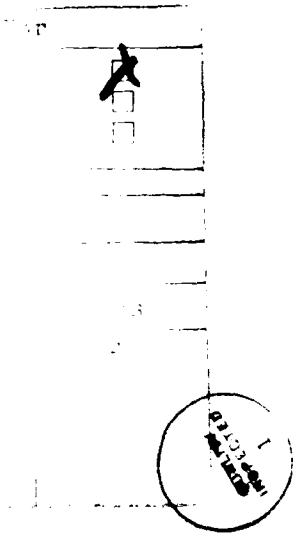
Recently we have also studied the interpretability of sequential neuronal messages. Although most neurophysiological experiments present images singly, images do not occur singly in normal vision. In light of our discovery that the response pattern carries stimulus-related information we have asked how the responses about sequentially presented stimuli can be interpreted. If responses to one stimulus affect messages arising in response to subsequent stimuli, then the messages could only be interpreted by accounting for this interaction. However, if the responses elicited by a stimulus are independent of the responses that preceded them, then interpretation of the neuronal messages is greatly simplified.

To learn how changes from one image to another affect stimulus-dependent neural messages, we recorded from primary visual cortex while Walsh patterns were were presented either singly for one of 18 durations or in pairs with varying intervals between them. Using our computational graphics facilities, we compared the results to two simple models. The first was one in which the response to the first stimulus would be terminated by the onset of the second, so that the response to the first stimulus would not affect the response to the second. The second model was a simple superposition model in which the response to the stimuli were added according to the appropriate timings.

When the time between sequentially presented stimuli was 66 msec or greater, the responses were independent, and both models (which under this condition predicted the same result) gave accurate predictions of the observed responses. At interstimulus intervals of 33 msec and less, the response to the first stimulus affected the start of the response to the second stimulus. The termination model gave poor predictions of the responses under this condition as the superposition model, which remained accurate. At interstimulus intervals of 0 and 16 msec the superposition model superposition failed. It appears that the initial peak of the response to the second stimulus was both decreased and delayed. However, at 33 msec or longer, the response was predicted by superposition. Based on these results we conclude that 30 msec is required to change temporally encoded messages so that each message can be interpreted without regard to the preceding message. This interval corresponds closely with the blank period that occurs between normal fixations. This result suggests that interpreting the messages transmitted by striate cortical neurons may be very straightforward. If the start of each message in a sequence emitted by striate cortical neurons can be recognized accurately, the responses can be interpreted without any reliance on knowledge of preceding messages.

Our results imply a new functional role for neurons in the visual system. Information about stimulus features is conveyed by individual neurons through multiple messages carried by a temporally modulated code. Like neurons in other parts of the visual system, lateral geniculate neurons simultaneously encode information about the luminance and luminance gradient of the scenes that fall on their receptive fields. Because the proportion of information transmitted by temporal modulation is substantially greater in LGN neurons than in ganglion cells, we hypothesize that one function of the LGN is to encode multiple stimulus features into a temporal code that keeps the information about different features separate.

Throughout the visual system, consolidation of local messages to determine global properties of images may be accomplished through compilation of many temporally encoded messages. Processing of information in visual areas may consist not so much in altering the distribution of active elements as in transforming temporally modulated messages. We suggest that the hierarchical organization of feature abstraction posited for the multiple visual areas should be replaced by a progression of spatial-to-temporal filtering that changes the emphasis of the visual features but never confounds or ignores information.



A-1

UNBIASED MEASURES OF NEURONAL INFORMATION TRANSMISSION AND CHANNEL CAPACITY

Lance M. Optican [1]

Timothy J. Gawne [2]

Barry J. Richmond [2]

Pinchas J. Joseph [1]

[1]Laboratory of Sensorimotor Research
National Eye Institute

[2]Laboratory of Neuropsychology
National Institute of Mental Health

This work was supported in part by
Air Force Office of Scientific Research Grant
AFOSR-ISSA-88-0073.

Address All Correspondence To:
Dr. L. M. Optican
Bldg. 10, Rm 10-C-101
National Eye Institute, NIH
Bethesda, MD 20892

phone: 301-496-3549

Draft
M.S. for
"Biological
Cybernetics"

June 16, 1989

CONTENTS

1. ABSTRACT

Biases are introduced into information measures based on neurophysiological data by 1) response quantization, 2) noise, and 3) small sample sizes. New measures of information transmission and channel capacity that minimize these biases are developed here. The properties of these new measures are demonstrated with data from simulations, and with data from individual neurons recorded from an awake monkey. The reduced bias of these new measures allows results from different analyses, or even different experiments, to be compared with confidence.

2. INTRODUCTION

Brain functions emerge from the rich interactions among individual neurons. These interactions depend on mechanisms that encode, process and transmit information from one area of the brain to another. Neurophysiological methods observe the consequences of this information processing, but have not yet provided an understanding of its underlying mechanisms. Information theory can provide a foundation for such an understanding by quantifying the encoding and transmission of information by neurons. However, its use in the study of neuronal mechanisms has been hindered because current methods for measuring information from experimental data yield upwardly biased estimates. This paper develops a new method for the unbiased estimation of transmitted information and channel capacity from neurophysiological data.

Information theory can be applied to neuronal systems if a neuron is considered as a channel linking a stimulus with a response. For convenience, it is assumed that the stimuli and the responses can be treated as discrete and finite sets of input and output symbols (also called sets of events or messages). Thus, for any experiment there is a discrete set of probabilities linking every stimulus event with every response event. The matrix whose j,k-th element is the conditional probability of receiving response r_k given that stimulus s_j was sent, $p(r_k \mid s_j)$, is called the *channel transition matrix*. The set of input and output symbols, the probabilities of the input symbols, and the channel transition matrix, completely specify the neuron's performance as a channel for transmitting messages.

Two information measures are commonly used to study neurophysiological systems: transmitted information and channel capacity. Since the order of stimulus presentation is randomized in the experiments, there is some uncertainty about which stimulus will be shown at any given time. After it is shown, though, the response of the neuron should indicate which stimulus was presented. However, more than one stimuli may give rise to the same response, and that response may be contaminated by noise. Thus, even after the response is received, some uncertainty about which stimulus was presented may remain. *Transmitted information* measures how much the uncertainty about which stimulus was presented is reduced by receiving its response. In a given experiment, the probabilities of the individual stimuli being presented are under control of the experimenter. Since all of the stimuli may not be equally differentiable by the neuron, changing the input probabilities would alter the amount of information transmitted by the neuron in the experiment. *Channel capacity* measures the maximum amount of information that the neuron could transmit, under the given experimental conditions, but with any set of input probabilities.

The application of information theory to neuronal data suffers from the same difficulties encountered with other biological data. The data are usually measured on a continuous scale, whereas the discrete form of information theory is most often used (the quantization problem). Biological systems show highly variable responses (the noise problem). Also, experimental considerations often limit the number of replications per condition (the small sample size problem). These difficulties prevent the accurate estimation of the probabilities needed to calculate information.

The measurement of information in experimental data is usually done with contingency tables (discrete input/output histograms). The conversion from continuous to discrete response variables allows the use of discrete probabilities in the information calculations. However, information estimated from contingency tables is known to be biased upward. It has been

shown that this is an additive bias, which may be subtracted off to obtain an unbiased estimate (Carlton 1969; Fagen 1978). There is an exact equation for this bias, but it requires complete knowledge of all the stimulus-response probabilities (Carlton 1969). Unfortunately, these probabilities are usually not known. Several approximations have been developed, but these either overstate the bias for small sample sizes (Carlton 1969; Miller 1955; Macrae 1971), or require assumptions about the distribution of probabilities underlying the process (Macrae 1971).

Contingency tables are not even an effective method for converting continuous variables into discrete variables. If a continuous variable is quantized by simply assigning it to the nearest bin, the information will increase logarithmically as the bin width decreases (Gallager 1968). Sakitt proposed a method of forming contingency tables by quantizing responses according to their rank order (Sakitt 1980; Sakitt et al. 1983). However, this technique is severely flawed: it can not be applied to multidimensional data (Optican and Richmond 1987), and it biases the information upward at the low and high ends of the data range (Crowe et al. 1988).

Previous attempts to measure information have considered some of these problems. Fagen corrected the sample bias (the error resulting from small sample sizes) in one-dimensional tables by two methods (Fagen 1978). In one method, Fagen subtracted an estimate of the sample bias obtained with the jack-knife statistical technique. In the second method, an estimate of the sample bias obtained from first-order approximation formulae was subtracted. Fagen's methods do not deal with the quantization problem, nor do the closed-form approximations seem applicable to neurophysiological data, where the underlying distributions are not known.

Optican and Richmond used a non-parametric method of forming the contingency table to minimize the quantization error, but did not correct for sample bias (Optican and Richmond 1987). Their technique for calculating information was based on an approach using a kernel estimate of the joint stimulus-response probability density function (Optican and Richmond 1987; Fukunaga 1972). This method eliminated the bias caused by bin-edge artifacts. However, we have since determined that it is still extremely sensitive to the number of data points in the sample, which leads to an overestimate of transmitted information. Also, their computer program for calculating information assumed, incorrectly, that the per stimulus covariance matrix was diagonal. This resulted in a small underestimation of transmitted information.

Thus, none of the available methods for calculating information measures accounts for all three bias sources presented above. This paper develops a new method for estimating transmitted information and channel capacity from multidimensional data, and tests it with simulated data sets with known distributions. The new method is then applied to data from individual neurons in primate visual cortex.

3. THEORETICAL DEVELOPMENT

An experiment can be considered as a set of stimulus-response pairings. The data determine the probability of occurrence of any stimulus-response pair. Calculating information measures from experimental data depends on estimating the joint stimulus-response probability density function (pdf), $p(s_j, r_k)$. Our new method of information analysis of neuronal activity requires the combination of several methodologies. First, the stimulus-response records must be converted to stimulus-response symbol pairs. Second, a channel probability matrix for the neuron must be estimated from the symbol pair data. Third, the sample-size bias and the signal-to-noise ratio must be estimated. Fourth, the amount of information transmitted by the neuron, and its capacity, must be estimated.

3.1. Information Theory

Information theory was developed over about thirty years, culminating with Shannon's definitive work in quantifying the way reliable information could be transmitted through noisy communication channels (Shannon 1948; Abramson 1963; Blahut 1987). If we regard the neuron as a channel for transmitting information about a stimulus, we can apply Shannon's information theory to the analysis of neural data. This will enable us to quantify the stimulus-dependent information transmitted by the neuron. Information theory deals strictly with the probabilities of combinations of input and output symbols, so it can quantify the stimulus-response relationship of the neuron without making any assumptions about the mechanisms involved (e.g., linear or nonlinear).

To quantify a neuron's response, some measure of its activity must be used. The mean firing rate, or number of action potentials (spikes) in a window (the spike count) are often used to give a univariate measure of the response. However, the response of neurons is clearly multivariate (Richmond and Optican 1987). To avoid making assumptions about what aspect of the neuron's response is used to encode messages about a stimulus, a complete representation of the response is sought. The time required to process the data goes up rapidly with dimensionality, so it is also necessary to find a compact representation. We have used the principal components of a continuous representation of neuronal activity as a multidimensional quantification of the neuronal response (Richmond and Optican 1987). A kernel estimate of the joint stimulus-response pdf is then formed, thus avoiding edge-effects in quantizing the response. This quantization is then used to form a discrete, extended code for the response (Optican and Richmond 1987).

3.2. Code Selection

To make calculations in information theory, events or messages must be encoded by symbols. Information theory does not specify how to assign symbols to the events. Thus, a specific code must be chosen for the channel. In our studies, the stimuli are chosen from a complete set of orthogonal patterns. Thus, it is easy to assign a code symbol to each input pattern. For example, for the 64 black and white Walsh patterns, ones with positive contrast are numbered 0-63, while those with negative contrast are numbered 64-127. This assignment of symbols to the stimuli is natural, since each pattern is orthogonal to all the others and may be considered an independent picture.

If stimuli varying across more than one dimension are used, care must be taken in the code assignments. For example, suppose the stimuli vary in two parameters: form and

brightness. If there were 16 forms and three levels of brightness, every combination of form and brightness would have to be in the stimulus set for completeness. Thus, an input code with 16 * 3 = 48 elements would be required. Such a code could be considered as a compound code with two "letters", one for the form and one for the brightness.

The selection of the output code is very difficult because the intrinsic code used by neurons is not known. It is generally accepted that the timing of the action potentials is the carrier of neuronal information. However, how the signal is encoded by that timing is not known. One solution to this problem is to assume different codes, and then to compare the amount of stimulus-dependent information conveyed by each code.

The most commonly used code is the mean-firing-rate, or spike count code. In this code, the average activity of the neuron in some interval is a continuous variable representing the response. The discrete code symbol is obtained by quantizing this continuous variable. Another code that has been used for neural firing is the binary code formed by dividing a response interval into small bins, and assigning a 1 or a 0 to each bin depending upon the presence or absence of a spike in that bin (Eckhorn and Pöpel 1974). The spike-count code is too restrictive because it assumes that only the number of spikes in the interval, and not their distribution, is important. The binary code is too inclusive, placing great emphasis on the exact timing of each spike. Furthermore, the spike-count code requires the specification of a response interval, and the binary code requires the specification of both a response interval and a bin width. These specifications are completely arbitrary.

A third suggestion for a neuronal code uses the statistical properties of the responses to obtain an intrinsic definition of the time scale of the code (Optican and Richmond 1987). This intrinsic time scale is obtained by extracting the principal components of the response waveforms from the kernel-estimated spike-density function. These principal components form an orthonormal basis set for any temporal waveform (Ahmed and Rao 1975). Since the time domain, or window, of the principal components are determined from the variations in the responses themselves, it does not depend upon arbitrary selection of an interval. Also, no arbitrary selection of bin width is needed, since the dimension of the principal components can be determined from the rank of the data domain covariance matrix, or from Nyquist's sampling theorem if the signal's bandwidth is known.

Previous work on primate neurons has shown that several of these principal component's coefficients are modulated by the stimulus, thereby requiring a multivariate measure of the neuron's response (Richmond and Optican 1987). The coefficient of each principal component is quantized to obtain a discrete set of multidimensional code symbols. An extended code that represents the multidimensional characteristics of the neuron's activity can be formed by concatenating the code symbols for the individual principal components. The information transmitted by the spike-count code with that transmitted by temporal waveform codes based on different numbers of principal components can then be compared to see which code is more useful.

3.3. Entropy

The symbols in a code are treated as events which can occur with a certain probability. For example, if a channel can transmit two symbols, say 0 and 1, and the probability of either symbol being sent is one-half, then before a symbol is sent there is a 50-50 chance it could be a 0 or a 1. The uncertainty about which symbol might be selected is very important in

information theory, and is given a special name: *entropy*. The entropy, H, is the average *a priori* uncertainty about an event:

$$H(S) = -\sum_j p(s_j) \log p(s_j). \quad (1)$$

Where S is the set of input symbols, $\{ s_j \}$. The individual symbols s_j have probability of occurrence $p(s_j)$. Note that the entropy is calculated from the probabilities of events, and does not involve any knowledge of the process that generated the events. This makes information measures *model-free*. Note that the measure of entropy is logarithmic. The choice of the base of the logarithm determines the units of information. The base two logarithm used throughout this paper yields information in *bits*.

3.4. Equivocation

If the entropy is the uncertainty about which symbol will be sent next, what is the uncertainty about which input symbol was sent after an output symbol has been received? If the channel is noiseless, there is no remaining uncertainty. However, if the channel is noisy, then there will be some probability of the symbol being received incorrectly. Nonetheless, even in the presence of noise, it should be possible to make a better guess after receiving the output of the channel than before. The uncertainty about which symbol was sent that remains *after* receipt of a symbol is called the *channel equivocation*:

$$H(S|R) = -\sum_k p(r_k) \sum_j p(s_j | r_k) \log p(s_j | r_k). \quad (2)$$

R is the set of output symbols, with individual symbols r_k. S is the set of input symbols, with individual symbols s_j. The backward channel conditional probability, that stimulus symbol s_j had been sent given that response symbol r_k was received, is $p(s_j | r_k)$.

3.5. Transmitted Information

If the equivocation is less than the entropy, then some information is said to be transmitted by the channel. This reduction in uncertainty, the difference between the entropy and the equivocation, is called the *gain in information*, the *mutual information*, or the *transmitted information* (Abramson 1963; Kullback 1959):

$$I(S;R) = H(S) - H(S|R) \quad (3)$$

Mathematically, the mutual information is symmetric, that is:

$$I(S;R) = I(R;S) \quad (4)$$

However, when considering a neuron as the channel, it is most meaningful to regard the neuron's response as transmitting information about the stimulus. Hence, we use the

transmitted information terminology, and use the symbol T instead of I (Eckhorn and Pöpel 1974). Thus, the average information about the stimulus (S) transmitted by the neuron in its response (R) is:

$$T(S;R) = H(S) - H(S|R) \tag{5}$$

3.5.1. Numerical Example

Transmitted information can be calculated by knowing the probabilities of stimulus-response pairs. If an experimental design is complete, the data can be used to determine this joint pdf. There were 128 stimuli in our experiment. If they all occurred equally often, the marginal probability of any particular stimulus occurring would be $p(s_j) = 1/128$. The entropy of such a *symbol source* would be 7 bits (from Eq. 1, since $\log_2(128) = \log_2(2^7) = 7$). Thus, the uncertainty about which symbol may be sent next by the source is 7 bits. If the symbol sent and the symbol received were statistically independent, then one would still have 7 bits of uncertainty left after receiving the signal. Alternatively, if there were a noiseless relation between the signal and the symbol sent, then there would be no uncertainty, or 0 bits, left after receipt of the signal. Between these two extremes, there would be a probabilistic relationship between symbol and signal, and the signal would tell something about the symbol sent, with some residual uncertainty. For example, the response might indicate only that the stimulus could have been any one of, say, eight patterns. This would leave a residual uncertainty of 3 bits ($\log 8 = 3$). Hence the information gain from receipt of the symbol would be $(7 - 3) = 4$ bits.

3.6. Channel Capacity

The transmitted information should be regarded as a function of both the channel itself, and the way the channel was used in the experiment (Blahut 1987):

$$T(S;R) = T(p(s_j); p(r_k | s_j)) \tag{6}$$

The way the channel is used is controlled by setting the *a priori* probabilities of stimulus occurrence, $p(s_j)$. The *channel capacity* is a measure of the maximum amount of information which the channel is capable of transmitting with any *a priori* input distribution:

$$C = \max_{p(s_j)} T(p(s_j); p(r_k | s_j)) \tag{7}$$

Note that when applied to neurons, the capacity is not necessarily the absolute maximum amount of information that neuron is capable of transmitting. Rather, it is the maximum amount it could transmit using the stimuli in the present experiment. If the stimuli were changed, it might be possible to alter the channel matrix of the neuron, and get more information through. For example, if stationary stimuli are used to study motion-sensitive neurons, then part of the neuronal channel is being ignored.

3.7. Joint Stimulus-Response Probability Density Function

Two things are needed to calculate anything in information theory: the sets of stimulus and response code symbols, and the probabilities of occurrence of those symbols. During an experiment the presentation of the stimuli are controlled and the corresponding responses measured. Thus, an experiment provides the data needed to calculate the joint probability of any stimulus-response pair, $p(s_j, r_k)$, since it is just the probability that a given stimulus and a given response occurred together in the experiment. There are two steps to estimating $p(s_j, r_k)$. First, all the stimuli and all the responses are converted to discrete codes. Then, a histogram is made that averages all the stimulus-response code pairs. After all of the data points have been averaged the histogram is a discrete estimate of $p(s_j, r_k)$.

Suppose there are N trials in the experiment, and $h(j,k)$ is a two-dimensional array where j and k range over the number of codes. For the spike-count code the response set includes M symbols, where M is the number of bins in the code histogram. For a q-dimensional temporal code based on q principal components, the number of members of the response set is M^q (where the multiple dimensions are treated as an extended code) (Abramson 1963; Optican and Richmond 1987). Let n be the number of trials wherein stimulus s_j elicited response r_k. Then the value in the j,k-th bin of the histogram, h(j,k), is just the fraction n/N. Information measures are often calculated based on this histogram, called a contingency table, since when N is large the histogram closely approximates the joint pdf, $p(s_j, r_k)$. However, if N is small, then the measure of information based on the contingency table is biased upward (Fagen 1978).

One cause of the bias in contingency tables is the quantization artifact. This artifact arises when a continuous representation of the response (either spike-count or coefficients of principal components) is forced into one of the M bins of the histogram. In other words, a response very near the edge of one bin is nonetheless assigned to the center of the bin. One way to reduce this quantization artifact is to form a continuous estimate of the joint pdf, $p(s_j, r_k)$, and then quantize that estimate (Optican and Richmond 1987). The amount of information calculated will depend on M. For neuronal data, M = 14 gives reasonable amounts of information independent of the number of response dimensions (Figure 1).

One method for forming a continuous estimate of the joint pdf that can be extended to multivariate response measures is to use kernel estimation (Fukunaga 1972; Silverman 1986). In kernel estimation, each data point is replaced with a continuous density function, centered on that point. The kernel estimate is the average of all these density functions.

Fukunaga has suggested that a good kernel for such an estimator is a Gaussian pulse with the same variance as the distribution of the data points themselves. This kernel is formed by using the sample covariance matrix of the data, Σ_D, as the covariance matrix of a Gaussian function. Such a kernel has the advantage that the statistical properties of the data, up to the second moment, are taken into account (Fukunaga 1972).

The estimate of the multivariate pdf (for a single stimulus) based on n data points is:

$$p_n(s_j,\mathbf{r}) = \frac{1}{n}\sum_{k=0}^{n-1} g(\mathbf{r},\mathbf{r}_k,h) \tag{8}$$

where $\mathbf{r}$ is the multidimensional response variable, $\mathbf{r}_k$ is the k-th multivariate data value, and h

is a function of n:

$$h(n) = n^{-0.49/k} \tag{9}$$

where k is the number of dimensions. The multivariate Gaussian kernel function g() is:

$$(2\pi)^{-k/2} h^{-k} |\Sigma|^{-1/2} \exp[-\frac{1}{2} h^{-2} (\mathbf{r} - \mathbf{r}_k)' \Sigma^{-1} (\mathbf{r} - \mathbf{r}_k)] \tag{10}$$

where $|\Sigma|$ is the determinant of the covariance matrix, and $\mathbf{r}'$ is the vector transpose of $\mathbf{r}$. In our previous work, we suggested using as the value of Σ the principal component transform domain covariance matrix, Σ_T (Optican and Richmond 1987). The advantage of doing this is that only the diagonal of Σ_T is nonzero, and thus the multivariate kernel g() is separable. This separability leads to a great simplification in calculating the continuous estimate of the joint pdf, since each dimension of the response vector can be computed independently, and then the multivariate result can be found just by multiplying all the one-dimensional values together. However, this approach has the disadvantage that all the subpopulations, i.e., the responses to a single stimulus, are treated as having distributions of the same shape (that of the total population). This leads to a somewhat conservative estimator, since it overestimates the dispersion in the data of any stimulus subpopulation and thus underestimates the transmitted information.

A more accurate estimator can be constructed by using a different kernel for each stimulus subpopulation. The value of Σ used is just the covariance matrix of the multidimensional response vectors elicited by that stimulus:

$$\Sigma = \frac{1}{n} \sum_{k=0}^{n-1} (\mathbf{r}_k - \bar{\mathbf{r}})(\mathbf{r}_k - \bar{\mathbf{r}})' \tag{11}$$

where $\bar{\mathbf{r}}$ is the average of all n responses to the given stimulus.

In the temporal modulation code studied here, the response vector is made up of the coefficients of the principal components. In general, the Σ matrix obtained for one stimulus is not diagonal (although the average of all the Σ matrices must be diagonal because that corresponds to Σ_T). Thus the multivariate kernel for one stimulus is not separable and can not be calculated by multiplying a set of one-dimensional values together.

3.8. Multidimensional Density Estimation with Non-Separable Kernel

Solving analytically for the amount of a non-separable kernel that goes into each bin of the joint stimulus-response histogram is extremely difficult. However, these non-separable kernels can be generated easily by combining a linear transformation of the data with a Monte Carlo, or simulation, technique for estimating probability density functions. Basically, the procedure obtains the desired pdf by 1) transforming to a domain where the distribution has an easily generated form, 2) generating an appropriately distributed set of points, 3) transforming the cloud of points back to the original data domain, and 4) building a histogram

of the points from the transformed cloud.

First, a set of discrete points with a Gaussian distribution is obtained from a pseudo-random number generator (Bratley et al. 1987; Press et al. 1988). The output of the generator is a univariate variable with zero mean and unit variance. One sample is generated for each dimension of the response (e.g., one, two, three, etc.), and these are used as the components of a vector for each point. This set of points forms a "cloud" in multidimensional space, whose covariance matrix is diagonal (i.e., the distribution is separable). Second, the Karhunen-Loève Transform (KLT) of the data is obtained since the KLT of the data has a diagonal covariance matrix in the transform domain (Ahmed and Rao 1975). Third, the standard kernel formed by the cloud of separable points distributed with unit-variance along each axis is converted to a cloud of points with the appropriate variance. The variance is adjusted by multiplying each component of the cloud point's vector by the standard deviation of its axis (i.e., the square root of the diagonal element of the diagonalized covariance matrix). Finally, this new cloud is transformed back into the the original domain by the inverse KLT.

The number of points used in the cloud is not fixed, but increases rapidly with the number of dimensions in the response space. This increase is necessary because the region near the mean occupies a proportionately smaller share of the space's volume as the dimensionality of a distribution increases (Silverman 1986). In our case, the number of points in the cloud (see Table I) was determined empirically to do a reasonable job of estimating the transmitted information for multidimensional responses (see Figure 2).

When used in this way, the cloud of transformed points can be thought of as a discrete kernel. To form the kernel estimate of the joint stimulus-response pdf, each data point is replaced with this cloud. The number of cloud points in each bin is then counted. Finally, the histograms are normalized by dividing by the number of cloud points and the number of responses. If the number of points in the cloud is large enough, the values in the bins now closely approximate the desired joint probability density function.

3.9. Transmitted Information

Information measures all depend upon probabilities that can be derived from the joint pdf, $p(s_j, r_k)$. The marginal probabilities are:

$$p(s_j) = \sum_k p(s_j, r_k) \tag{12}$$

and

$$p(r_k) = \sum_j p(s_j, r_k) \tag{13}$$

The probability that a certain response occurs after a given stimulus is presented is called the conditional probability, $p(r_k \mid s_j)$. The conditional probability may be calculated using Bayes' law:

$$p(r_k \mid s_j) = \frac{p(s_j, r_k)}{p(s_j)} \tag{14}$$

The amount of information transmitted by the neuron about the particular stimulus s_j, averaged over all the responses in the set R, is the *conditional transmitted information*, $T(s_j ; R)$:

$$T(s_j ; R) = \sum_k p(r_k \mid s_j) \log \frac{p(r_k \mid s_j)}{p(r_k)} \tag{15}$$

where $p(r_k \mid s_j)$ is the conditional probability of getting response r_k given stimulus s_j. The summation is over all the members of the set R.

The *average transmitted information* can then be calculated using Eq. 5, or as the weighted sum of all the conditional transmitted informations:

$$T(S ; R) = \sum_j p(s_j)\, T(s_j ; R). \tag{16}$$

Since what we can actually calculate based on the experimental data is the joint probability density function, $p(s_j, r_k)$, it is useful to write another set of equations for the transmitted information. The conditional transmitted information (i.e., the transmitted information per code symbol) can also be expressed using Bayes' law (Eq. 14) as:

$$T(s_j ; R) = \sum_k \frac{p(s_j, r_k)}{p(s_j)} \log \frac{p(s_j, r_k)}{p(s_j)\, p(r_k)} \tag{17}$$

The average transmitted information can then be expressed as:

$$T(S;R) = \sum_j \sum_k p(s_j, r_k) \log \frac{p(s_j, r_k)}{p(s_j)\, p(r_k)} \tag{18}$$

3.10. Channel Capacity

The channel capacity can be calculated by an iterative procedure once the channel matrix, $p(r_k \mid s_j)$, is known (Arimoto 1972; Blahut 1972). The iterative procedure finds the set of stimulus probabilities, $\{p(s_j)\}$, that maximizes the information transmitted by the given channel. Blahut proves the following theorem and gives an iterative algorithm for its implementation: For any set of input probability distributions, $\mathbf{p} = \{p(s_j)\}$, define the function $c_j(\mathbf{p})$ as an exponential measure of the contribution of the j-th stimulus to the capacity (i.e., the conditional transmitted information for the j-th stimulus):

$$c_j(\mathbf{p}) = \exp[\sum_k p(r_k \mid s_j) \log \frac{p(r_k \mid s_j)}{\sum_i p(s_i)\, p(r_k \mid s_i)}] \tag{19}$$

If $\mathbf{p}^{(0)}$ is any strictly positive *a priori* stimulus distribution, then the series $\mathbf{p}^{(0)}$, $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, ..., defined by

$$p(s_j)^{(r+1)} = p(s_j)^{(r)} \frac{c_j(\mathbf{p}^{(r)})}{\sum_i p(s_i)^{(r)} c_i(\mathbf{p}^{(r)})} \tag{20}$$

is such that $T(\mathbf{p}^{(r)}; p(r_k \mid s_j))$ converges to C from below (Blahut 1972; Blahut 1987).

The convergence criterion given by Blahut is based on the difference between lower and upper bounds on the capacity (Blahut 1987):

$$\log(\max_j c_j(\mathbf{p})) - \log(\sum_j p(s_j)\, c_j(\mathbf{p})) < \varepsilon \tag{21}$$

To avoid the time of computing the log functions, this test was converted to an equivalent form. First, convert the log to the natural base, and use the approximation:

$$\exp(\varepsilon) \approx 1 + \varepsilon\,, \quad \varepsilon \ll 1 \tag{22}$$

Then the test becomes:

$$\frac{\max_j c_j(\mathbf{p})}{\sum_j p(s_j)\, c_j(\mathbf{p})} < 1 + \varepsilon \tag{23}$$

For multidimensional output codes, or for large numbers of stimulus codes, the computation time for the capacity was very large (Table II). Two other alterations to the algorithm were made to reduce that time. The first step in the iterative loop used to calculate the capacity was altered. This step updates the probability distribution of the stimulus by multiplying by:

$$\zeta = \frac{c_j(\mathbf{p})}{\sum_i p(s_i)\, c_i(\mathbf{p})} \tag{24}$$

This ratio must approach one (unless $p(s_j)$ is zero). To accelerate the convergence, we used ζ^2 for the first five iterations, and ζ thereafter. To further reduce the time to obtain an answer, the convergence criterion, ε, was usually set to stop the iterations after the capacity

was within 1 - 3% of the final value.

3.11. Small-Sample Bias

All the information calculations proceed from the $p(s_j, r_k)$ matrix. However, these results are severely biased by a small-sample artifact. This bias exists even with one dimensional data, and thus is common to all information calculations that depend upon limited amounts of experimental data. To understand where this bias comes from, consider an example where there are L stimuli with N replications per stimulus and M bins in the $p(s_j, r_k)$ matrix. In other words, the continuous data variable is to be encoded by quantizing it into one of M bins. Now, suppose that the stimulus and response are completely unrelated. In other words, the response contains zero bits of information about the stimulus. The only distribution of N samples per stimulus in M bins that yields zero bits of information is the one that has the same number of responses to every stimulus in every bin. Suppose that the responses fall into K bins ($K \leq M$). Then each of the L stimuli must contribute N/K responses to each bin. If the stimulus-response relation is random, then as N increases the number actually found, by chance, in each of the bins does approach N/K. However, for small values of N, especially if $N < K$, all of the bins will not be filled evenly. This uneven filling will result in the apparent transmission of information. This non-zero value is a bias in the estimate of transmitted information formed from the $p(s_j, r_k)$ matrix. Below we develop an improved estimator of transmitted information that removes this small-sample bias.

An improved estimator for any statistic can be formed using a Monte Carlo simulation technique called the bootstrap (Efron 1982). The bootstrap is used to build resampled sets of the stimulus-response pairings, but without regard to which stimulus elicited which response. Suppose that the number of times this resampling is performed is N_b. Let T_i be the transmitted information calculated from the i-th set of resampled data. Then we form a bootstrapped estimate of the transmitted information under the assumption of no stimulus-response association as:

$$T_b = \frac{1}{N_b} \sum_{i=0}^{N_b - 1} T_i \tag{25}$$

Efron gives as the variance of this estimator (Efron 1982):

$$\text{var} = \frac{1}{N_b^2} \sum_{i=0}^{N_b - 1} (T_i - T_b)^2 \tag{26}$$

For our data, the number of bootstrap repetitions did not need to be large. There was very little improvement after averaging about ten resampled sets (Figure 3). To minimize processing time, the value of $N_b = 5$ was used routinely, which amounts to an error of only 1 - 2%. These sets of arbitrary stimulus-response pairings lead to zero information if the number of replicates is large. If the number of replicates is small, then there will be an apparent transmission of information, because of the inability to evenly represent every response type in every stimulus class (the only condition that yields zero transmitted information). The boot-strapped information can thus be used as an estimate of the small-

sample bias. An improved estimate of the transmitted information, then, would result if some function of the boot-strapped information (T_b) was subtracted from the calculated transmitted information (T). To determine a good function of T_b to use, we considered three desirable properties of the improved estimator ($\hat{T}$):

1. $\hat{T}$ should be asymptotically unbiased (i.e., approach the correct solution smoothly as the sample size increases).
2. $\hat{T}$ should be zero for pure noise for all sample sizes.
3. $\hat{T}$ should not overestimate the correct value for small sample sizes.

One way to insure that the new estimator has the correct properties is to notice that the correction term should depend on an estimation of both the sample and noise biases. In particular, the new estimator should approach the calculated value of T as the noise goes down or as the sample size goes up. Since T_b itself declines asymptotically as the sample size goes up, it is only necessary to incorporate a dependence on the signal-to-noise ratio of the data. Various measures of noise could be derived from the data set (e.g., the ratio of within-group variance to between-group variance), but these will depend upon some assumptions about the noise process (e.g., Gaussian distribution). This is undesirable, since one advantage of using information theory is that its measures are model-free. Fortunately, noise can be estimated from the information measures themselves. The ratio of the estimated bias (T_b) to the transmitted information (T) is a measure of how noisy the data are. If the data are pure noise, then T and T_b will be virtually the same. If there is no noise, then T_b will be zero. So, the ratio T_b/T can serve as a factor that is sensitive to the signal-to-noise ratio of the data.

The improved estimator we seek is then simply:

$$\hat{T} = T - \frac{T_b}{T} T_b \tag{27}$$

The weighted bootstrap term is an estimate of the bias in T from both noise and small sample size sources. The performance of the improved estimator is shown in Figure 4. This correction can be applied in another way, since the transmitted information can also be computed as the weighted average of the conditional transmitted information per stimulus (cf. Eqs. 15 and 16 or 17 and 18). Thus, T in Eq. 27 can be considered as either $T(S;R)$ or $T(s_j;R)$. The two approaches give almost identical results. We use the second, with $T(s_j;R)$ in Eq. 27, since that also provides the corrected conditional transmitted information per stimulus.

Subtracting the weighted bias term is equivalent to multiplying by a correction factor, α:

$$\hat{T} = \alpha T \tag{28}$$

where

$$\alpha = 1 - \left(\frac{T_b}{T}\right)^{\gamma} \tag{29}$$

with γ equal to two. This suggests a family of correction factors, determined by the value of γ. The smaller the value of γ, the more conservative the new estimator is for small sample sizes. In our data, the choice of two for γ gave well behaved results for sample sizes below ten (Figure 5).

It is easy to see that the new estimator has all of the desired properties listed above. First, T is itself an asymptotically unbiased estimator of the true transmitted information, since it is based on a kernel estimate of $p(s_j, r_k)$, and that estimate is asymptotically unbiased (Fukunaga 1972). As N increases, the probability of finding the shuffled responses evenly distributed across the response space approaches one. The information transmitted by such an even distribution is zero. Hence T_i, the information transmitted after each reshuffling of the N data points, tends to zero for large N. This means that T_b, the average value of T_i, also tends to zero as N increases. Hence $\hat{T}$, which is just the weighted difference of T and T_b, must also be an unbiased estimator of the transmitted information.

Second, $\hat{T}$ is zero for pure noise for all values of N, since if the stimulus-response relation is pure noise, then shuffling should have no effect on the distribution of the responses. Hence, T and T_b will be the same, and their difference will, on average, be zero.

Third, the direct calculation of information seriously overestimates the information transmitted when N is very small (cf. Fig. 4). This overestimate arises because with small N there is a high probability that data points will be assigned to unique bins. This uniqueness leads to poor estimates of $p(s_j, r_k)$, which result in large amounts of information. However, this same problem applies to T_i as well, so that the bootstrapped bias correction term will also be large. Thus, their difference will be reasonably well behaved. Indeed, $\hat{T}$ underestimates information for values of N below about 10, which is more conservative than the overestimates provided by T.

The average mutual information is never less than zero (Abramson 1963; Blahut 1987). If the transmitted information is low enough, or the data are noisy enough, it is even possible for the average mutual information, $\hat{T}(S;R)$, to be slightly negative. This occurs whenever the true data points are more dispersed in the response space than their shuffled counterparts. Nevertheless, the advantages of using the new estimator outweigh the minor disadvantage of having small negative estimates for very noisy data.

In an exactly parallel argument, an improved estimator of the capacity can also be formed by subtracting a weighted bias term:

$$\hat{C} = C - \frac{C_b}{C} C_b. \tag{30}$$

C_b is the bootstrapped estimate of C after resampling the data without regard to the original stimulus-response pairings. The improved performance of this new estimator is shown in Figure 6. Its dependence on the weight factor's exponent (γ) is virtually identical to that shown for $\hat{T}$ in Fig. 5.

4. APPLICATION TO NEURONAL DATA

The improved estimators developed above were applied to recordings from individual neurons (Optican and Richmond 1986). The data used here came from two complex cells in the primary visual cortex of a monkey performing a fixation task. The stimulus set consisted of 128 stationary, black and white two-dimensional pictures based on Walsh functions. Several codes were used in the information calculations: a response strength (spike-count) code, and temporal waveform codes of different dimensionalities (1-5). The temporal waveform codes were based on from one through five principal components of the response (Richmond and Optican 1987; Optican and Richmond 1987). The data from these neurons were selected because they transmitted relatively large amounts of information (about one bit) and included relatively large numbers of repetitions (over thirty) for each stimulus.

Transmitted information and channel capacity were calculated using the formulae given above. One important index, the ratio of information transmitted in a temporal code to that transmitted in a mean-rate code, was also calculated. The higher this ratio, the more information about the stimulus is being transmitted in the temporally modulated component of the neuron's response.

Figure 7 shows the performance of $\hat{T}$ calculated for the neuronal data from subsets of different sizes. The values for large sample sizes are an indication of the true information measures. The performance of $\hat{T}$ is very good for sample sizes as small as seven.

We have shown that the temporal modulation of a neuron's response carries more information about the stimulus than the strength of the response alone (Optican and Richmond 1987). This was demonstrated by showing that more information was carried by a *multidimensional temporal code* based on the principal components. However, these conclusions were reached using the biased estimate of transmitted information, T. Figure 8 shows that information transmitted by a temporal code also increases with the dimensionality of the code, when using the improved estimator, $\hat{T}$. Figure 9 shows that the ratio of transmitted information in the temporal code to that in the strength code increases with the number of components. This suggests that useful information is available in the temporal modulation of the neuron's responses (Optican and Richmond 1987). The results for channel capacity are similar (Figures 10 and 11).

5. DISCUSSION

New estimators of transmitted information and channel capacity were based on a quantized kernel estimate of probability, a ratio estimate of noise, and a bootstrap estimate of small-sample bias. Results from simulated data and neuronal data show that the new estimators overcome the upward bias effects on information measures from three sources: 1) response quantization, 2) noise, and 3) small sample sizes. The improved estimators thus allow results from different experiments, or from different analyses on data from the same experiments, to be compared.

Crowe *et al.* (1988) have suggested that for information theory to be useful in biological studies, a new method must be developed that eliminates the problems of quantization and sample-size bias in estimating the transmitted information. The improved estimators of transmitted information and channel capacity developed here satisfy those criteria for application to biological data.

These improved estimators should benefit any biological studies that need to quantify the amount of information transmitted in an experiment. This is especially important in neurophysiology, where it is standard practice to ascribe special significance to that stimulus which evokes the strongest response from a neuron. With the new tools presented here, it is now possible to avoid assumptions about the important parameter of a neuron's response by quantifying the relative efficiency of codes based on different measures of response activity (such as spike count versus temporal modulation). Furthermore, it is now possible to judge the functional importance of different stimuli by measuring the stimulus-dependent information actually transmitted by the neuron. Thus, these improved estimators make it possible to apply information theory to achieve an understanding of how neurons perform the encoding and transmission of information, the bases of all brain function.

6. ACKNOWLEDGEMENT

We would like to thank Dr. Richard E. Blahut of IBM Corp. for discussions on calculating information measures.

7. TABLES

Table I.

Number of Points *vs.* Dimensionality. The left column, n, is the number of dimensions. The middle column is the number of points needed to estimate a Gaussian pdf with a relative mean square error of 0.1 (Silverman 1986). The right column is the number of points determined empirically for estimating joint stimulus-response pdf.

n	Gaussian pdf	$p(s_j, r_k)$
1	4	100
2	19	500
3	67	2000
4	223	5000
5	768	8000

Table II.

Computer time *vs.* Dimensionality. Times to calculate information measures using a Silicon Graphics Iris 4D-120 computer (running at about 10 million instructions and 2 million floating point operations per second). The number of bootstraps was 5, the number of stimuli was 128, and the number of stimulus repetitions was 30. Number of dimensions, n; time (minutes) for transmitted information, $\hat{T}$; time (minutes) for channel capacity, $\hat{C}$.

n	$\hat{T}$	$\hat{C}$	
1		1	5
2		2	38
3		9	251
4		112	
5		331	

8. FIGURE CAPTIONS

Figure 1. Dependence of improved estimator of information, $\hat{T}$, on number of bins per response dimension in the $p(s_j, r_k)$ matrix. $\hat{T}$, for temporal codes with 1 - 5 principal components, was calculated using data from a single neuron recorded in primary visual cortex of a monkey during a fixation task (Optican and Richmond 1986). As the number of bins becomes smaller, the effects of noise are exaggerated. As the number of bins becomes larger, the effects of small sample are exaggerated. Thus, $\hat{T}$ shows a maximum (near 14), for higher dimensions.

Figure 2. Dependence of transmitted information, T, on the number of points in the Gaussian cloud used to estimate the joint stimulus-response pdf, $p(s_j, r_k)$. T was calculated for temporal codes with 1 - 5 principal components. For three or more dimensions, T is biased upwards for small numbers of cloud points. The dotted line (open circles) shows T calculated with number of cloud points suggested by Silverman for pure gaussian distributions (Silverman 1986). The dashed line (filled circles) shows T calculated with the number of cloud points chosen in this work.

Figure 3. Dependence of $\hat{T}$ on number of bootstrap resamplings. Error is at most 2 - 3%. Five resamplings were used for computing $\hat{T}$, a compromise between speed and accuracy good to within a few percent.

Figure 4. Dependence of transmitted information on sample size for simulated three-dimensional data. Data for panels A and B were generated by a three-dimensional stochastic process, where two components were independent gaussian-distributed noise, and the third component was independent uniformly-distributed noise. No information should have been transmitted by these data. Data for panels C and D were generated by adding a deterministic signal to the stochastic process just described. The first deterministic component conveyed two bits, the second conveyed zero bits, and the third conveyed one bit of information. The step size between deterministic code levels was five times the standard deviation of the noise. Because of the noise, slightly less than three bits of information should have been transmitted by these data.

100 data points were generated, and the information measures were calculated for smaller subsets of these points (different numbers of stimulus repetitions). In panels A and C, T (solid line) is the calculated transmitted information, B (dotted line) is the bias estimate, and T_b (dashed line) is the noise-factor weighted bias estimate. In panels B and D, T_d (dashed line) is an estimator of transmitted information based on the bias estimate, without the noise-factor weighting (the difference between T and B). $\hat{T}$ (solid line) is the improved estimator of transmitted information, with the noise-factor weighted bias estimate (the difference between T and T_b). The horizontal dotted line is the amount of information carried by the deterministic component of the signal. Note that for some sample sizes, $\hat{T}$ is negative (instead of zero) in panel B. In panel D, note that the effect of the noise-factor weighting on $\hat{T}$ is to accelerate its convergence to the correct value for much smaller sample sizes than T_d.

Figure 5. Dependence of transmitted information on the bias weight exponent, γ (Eq. 29). This figure is based on the same simulated data used in Fig. 4. In panel A, the solid line is the transmitted information, T. The dotted lines are a family of curves representing the weighted bias correction term, T_b, for different values of γ (1.0, 1.5, 2.0, 2.5, 3.0). As γ increases, the weighted bias term decays faster. Panel B shows a family of curves representing the improved estimator, $\hat{T}$, for these values of γ. As γ increases, $\hat{T}$ rises faster.

The dots lie on the curves with $\gamma = 2.0$, its nominal value. The horizontal dotted line is the amount of information carried by the deterministic component of the signal.

Figure 6. Dependence of capacity on sample size for simulated three-dimensional data. This figure is based on the same simulated data used in Fig. 4. Panels A and C show the capacity, C, bias, B_c, and the noise-factor weighted bias, C_b. Panels B and D show the the improved estimator, $\hat{C}$, and an estimator based on the bias alone, C_d. The horizontal dashed line is the capacity of the deterministic channel. Note that $\hat{C}$ can go negative for small values (panel B). Note that $\hat{C}$ converges to its correct value faster than C_d because of the noise-factor weighting (panel D).

Figure 7. Dependence of transmitted information on sample size (number of stimulus repetitions) for neuronal data. Transmitted information was calculated for temporal codes with 1 - 4 principal components. In panel A, the solid line is T and the dotted line is T_b. In panel B, the solid line is $\hat{T}$. For two or more dimensions, T is biased upwards for small sample sizes. However, T_b also rises for small sample sizes. The difference, $\hat{T}$, forms a conservative estimate of transmitted information, since it is too small for small sample sizes. Values of $\hat{T}$ are usable for much smaller sample sizes than those required by T.

Figure 8. Additivity of transmitted information as dimensionality of the temporal waveform code increases. Transmitted information was calculated from neuronal data using temporal codes with 1 - 5 principal components. $\hat{T}$ for the individual components is shown by the dashed line (filled squares, labeled **pc**). Because the principal components are ordered by the amount of variance in the signal they account for, this line must decrease for higher components. The dashed line (open circles, labeled **sum**) is the sum of $\hat{T}$ calculated for each component of the waveform code individually, and is an upper bound on the joint $\hat{T}$. The solid line (filled circles, labeled **joint**) is $\hat{T}$ calculated using codes of increasing dimensionality. Note that the total information appears to be approaching an asymptote below that of the summed components. This failure of additivity indicates that some of the information contained in each component is redundant. As a control, information was also calculated using the same principal component in each dimension. This corresponds to placing the data on a line along the hyperdiagonal of the response space. This dotted line (open diamonds, labeled **ctrl**) should be horizontal, since the amount of information is the same, even if the data are redistributed along a line in response space with a different orientation. However, the control line drops gradually with increasing dimensionality (up to 25% by five dimensions). This drop indicates that information is lost, presumably because the very large number of bins needed to quantize the higher dimensional response space exacerbates the small-sample size problem. This loss of information with higher dimensions makes $\hat{T}$ a conservative estimate (i.e. an underestimate) of multidimensional transmitted information.

Figure 9. The ratio of transmitted information conveyed by temporal waveform codes to that conveyed by a response strength code. Transmitted information was calculated from neuronal data using codes based on 1 - 5 principal components ($\hat{T}$), and using a code based on response strength (spike count, $\hat{T}_s$). The ratio of the waveform and strength codes indicates the amount of stimulus-dependent information present in the temporal modulation of the neuron's activity. The five component code conveyed over three times as much information as the univariate strength code.

Figure 10. Dependence of channel capacity on sample size (number of stimulus repetitions) for neuronal data. Channel capacity was calculated for temporal codes with 1 - 3 principal components. In panel A, the solid line is C and the dotted line is C_b. In panel B, the solid line is $\hat{C}$. C is biased upwards for small sample sizes. However, C_b also rises for small sample sizes. The difference, $\hat{C}$, forms a conservative estimate of channel capacity, since it is too small for small sample sizes. Values of $\hat{C}$ are usable for much smaller sample sizes than those required by C. This dependence on sample size is similar to that of transmitted information shown in Fig. 7.

Figure 11. The ratio of channel capacity using temporal waveform codes to that using a response strength code. Channel capacity was calculated from neuronal data using codes based on 1 - 3 principal components ($\hat{C}$), and using a code based on response strength (spike count, $\hat{C}_s$). The ratio of the waveform and strength codes indicates the relative capacity of the neuron considered as an information channel using temporal modulation of the neuron's activity as opposed to using only its strength. The three component code conveyed almost three times as much information as the univariate strength code.

9. REFERENCES

Abramson N (1963) Information theory and coding. McGraw-Hill, New York

Ahmed N, Rao KR (1975) Orthogonal Transforms for Digital Signal Processing. Spring-Verlag, Berlin

Arimoto S (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Info. Theory IT-18:14-20

Blahut RE (1972) Computation of channel capacity and rate-distortion functions. IEEE Trans. Info. Theory IT-18:460-473

Blahut RE (1987) Principles and Practice of Information Theory. Addison-Wesley, Reading, Mass.

Bratley P, Fox BL, Schrage LE (1987) Second Edition. Springer-Verlag, New York

Carlton AG (1969) On the bias of information estimates. Psychol. Bull. 71:108-109

Crowe A, de Ruiter T, Blaauw M, Oosthoek B (1988) Information transmission in non-visual fingertip matching along a horizontal track in the median plane. Biol. Cybern. 58:141-148

Eckhorn R, Pöpel B (1974) Rigorous and extended application of information theory to the afferent visual system of the cat. I. basic concepts. Kybernetic 16:191-200

Efron B (1982) The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia

Fagen RM (1978) Information measures: statistical confidence limits and inference. J. theor. Biol. 73:61-79

Fukunaga K (1972) Introduction to Statistical Pattern Recognition. Academic Press, New York

Gallager RG (1968) Information theory and reliable communication. Wiley, New York

Kullback S (1959) Information theory and statistics. Wiley, New York

Macrae AW (1971) On calculating unbiased information measures. Psychol. Bull. 75:270-277

Miller GA (1955) Note on the bias of information estimates. Information Theory in Psychology; Problems and Methods II-B:95-100

Optican LM, Richmond BJ (1986) Temporal encoding of pictures by striate neuronal spike trains: II. Predicting complex cell responses. Soc. Neurosci. Abstr. 12:

Optican LM, Richmond BJ (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. J. Neurophysiol. 57:162-178

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) The Art of Scientific Computing. Cambridge Univ. Press, Cambridge

Richmond BJ, Optican LM (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: II. Quantification of response waveform. J. Neurophysiol. 57:147-161

Sakitt B (1980) Visual-motor efficiency (VME) and the information transmitted in visual-motor tasks. Bull. Psychonom. Soc. 16:329-332

Sakitt B, Francis L, Zeffiro TA (1983) The information transmitted at final position in visually triggered forearm movements. Biol. Cybern. 46:111-118

Shannon CE (1948) A mathematical theory of communication. Bell Syst. Tech. J. 27:379-423

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London

Fig -

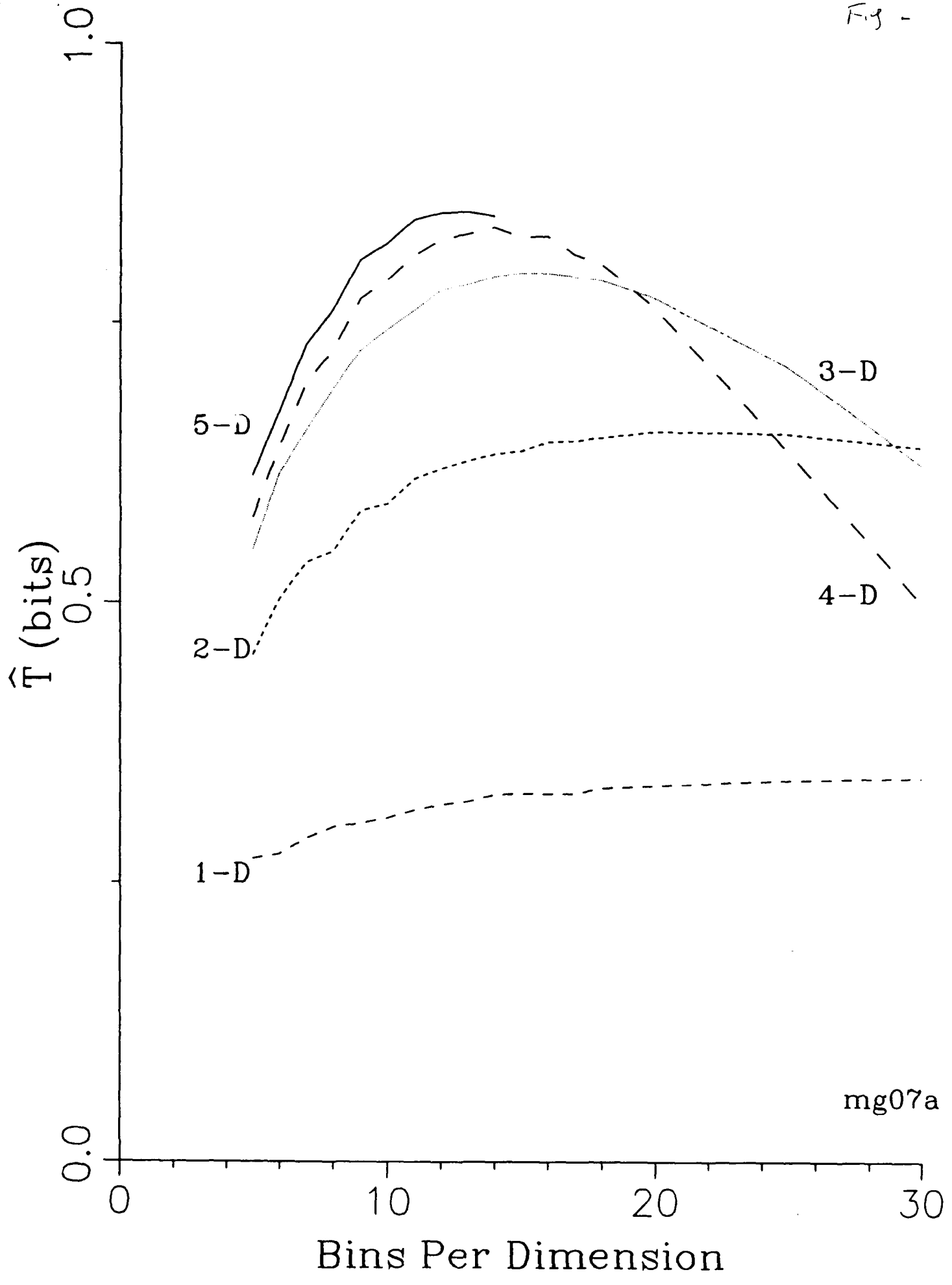


Fig. 2

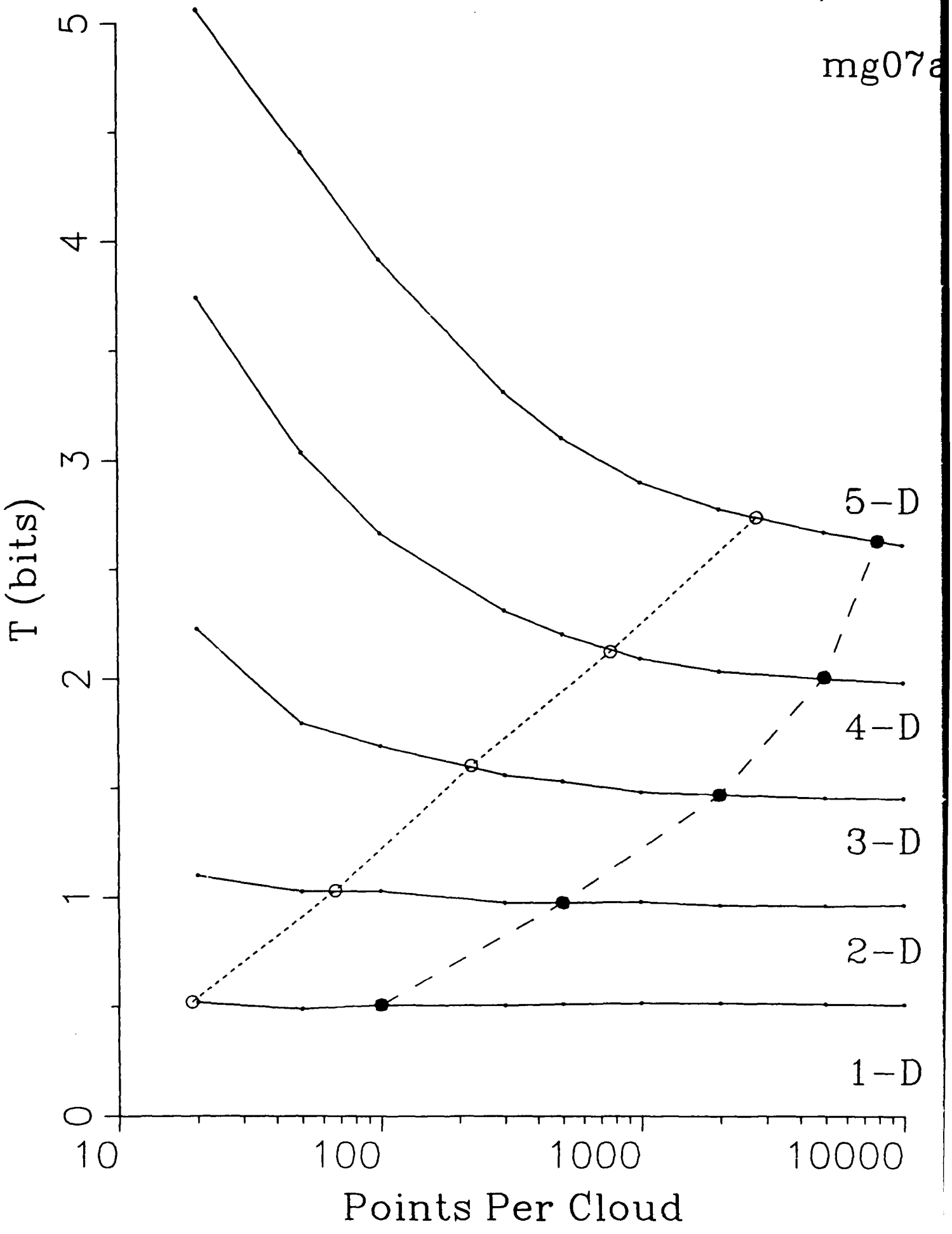
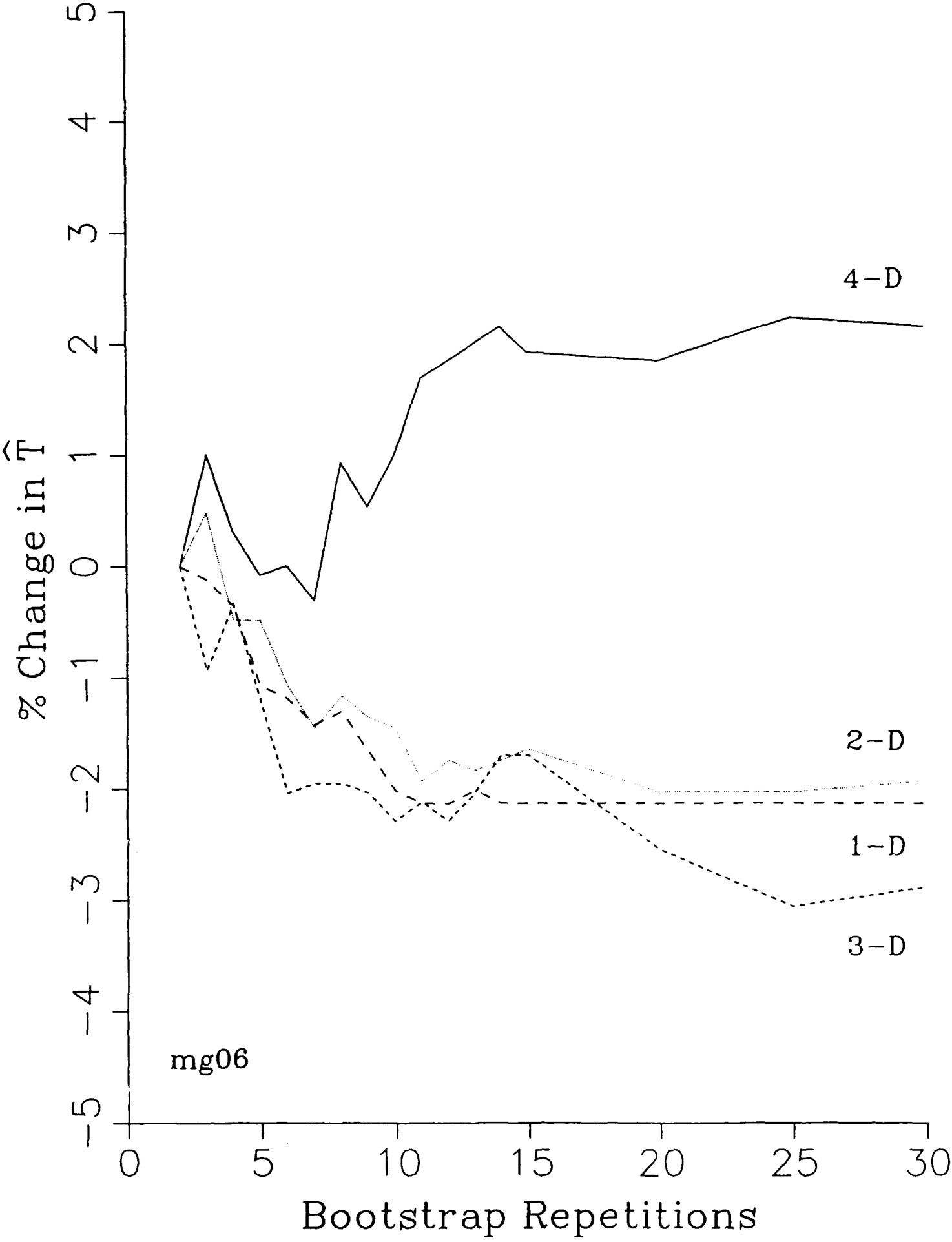
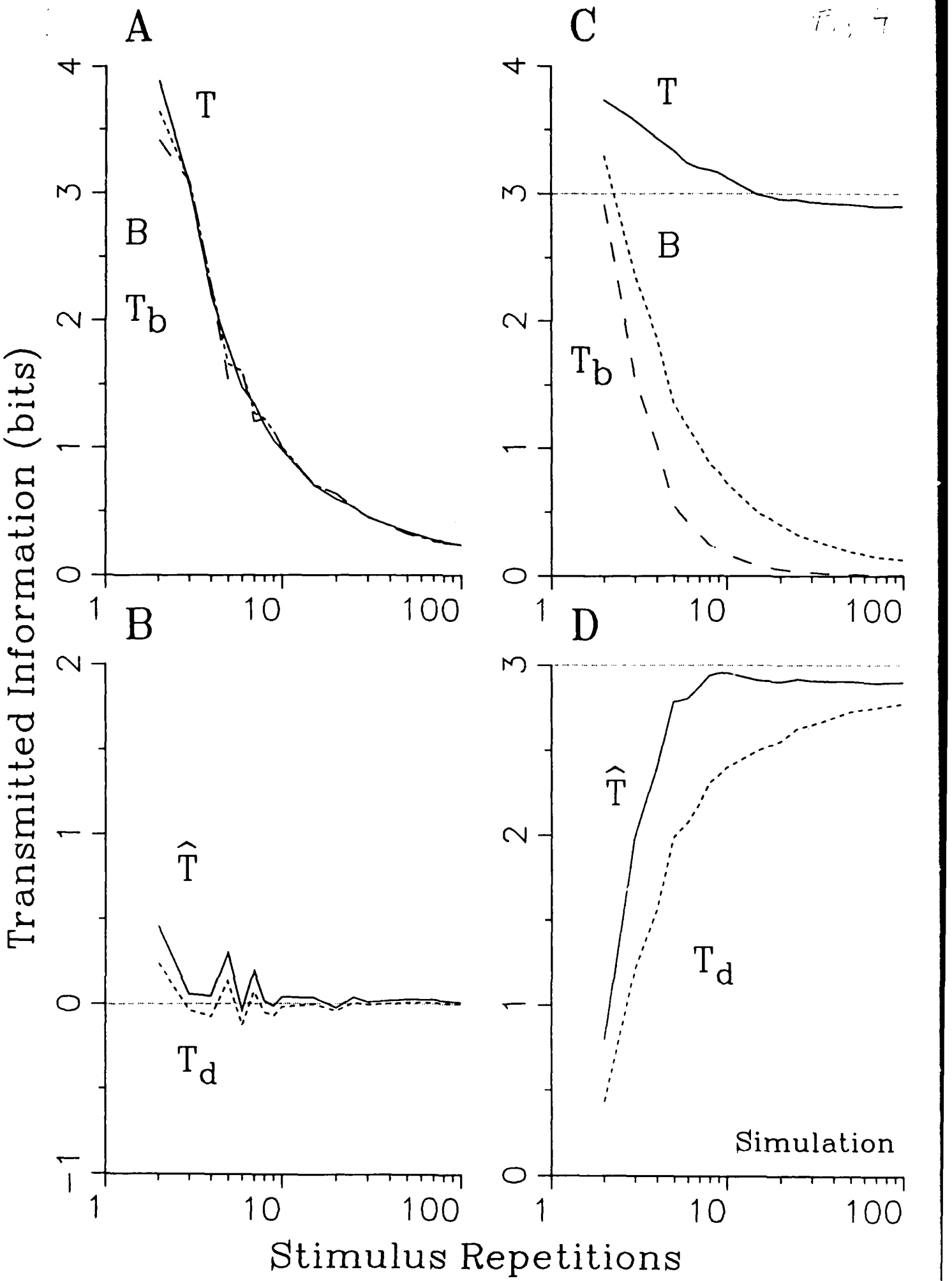
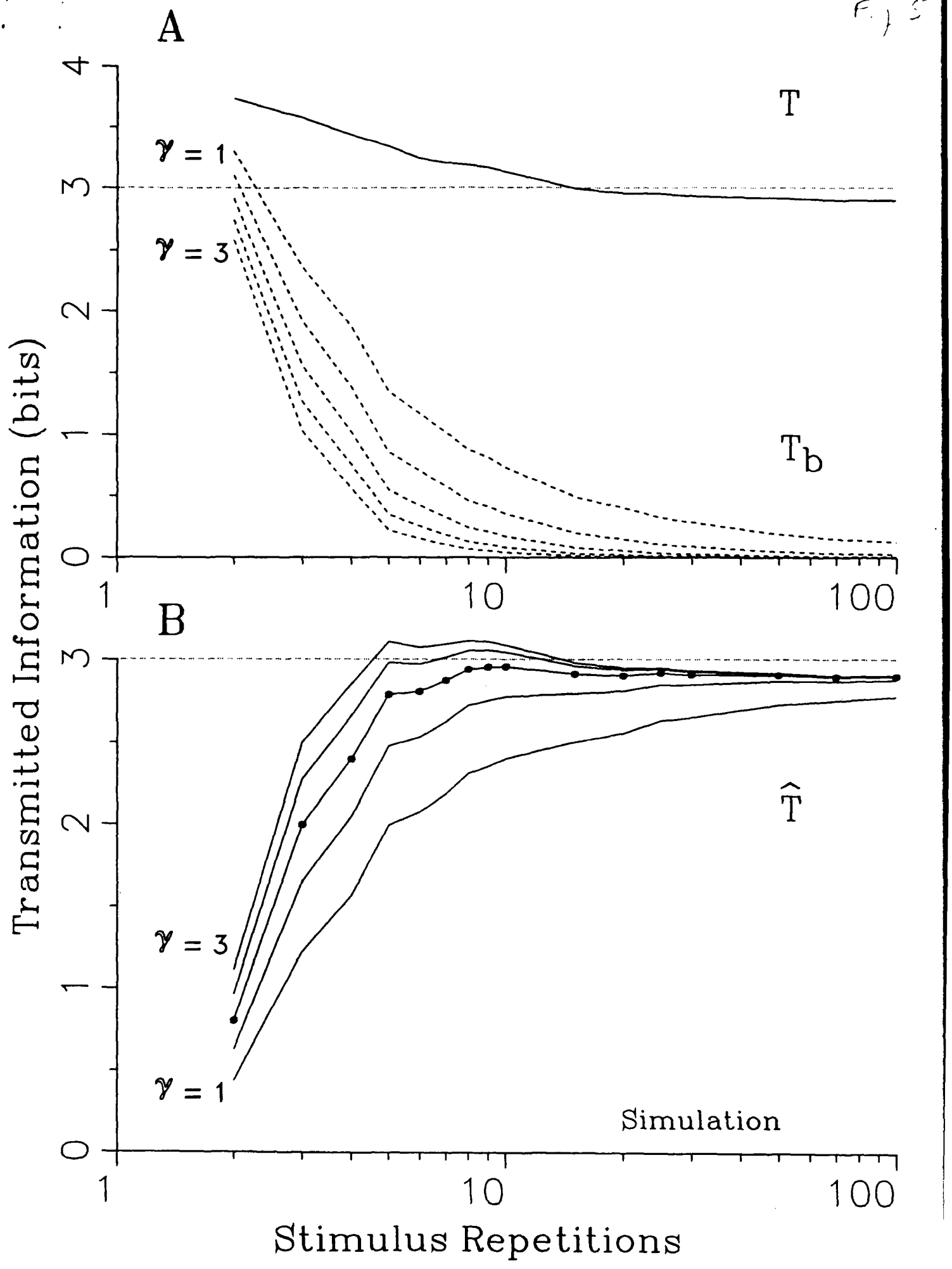
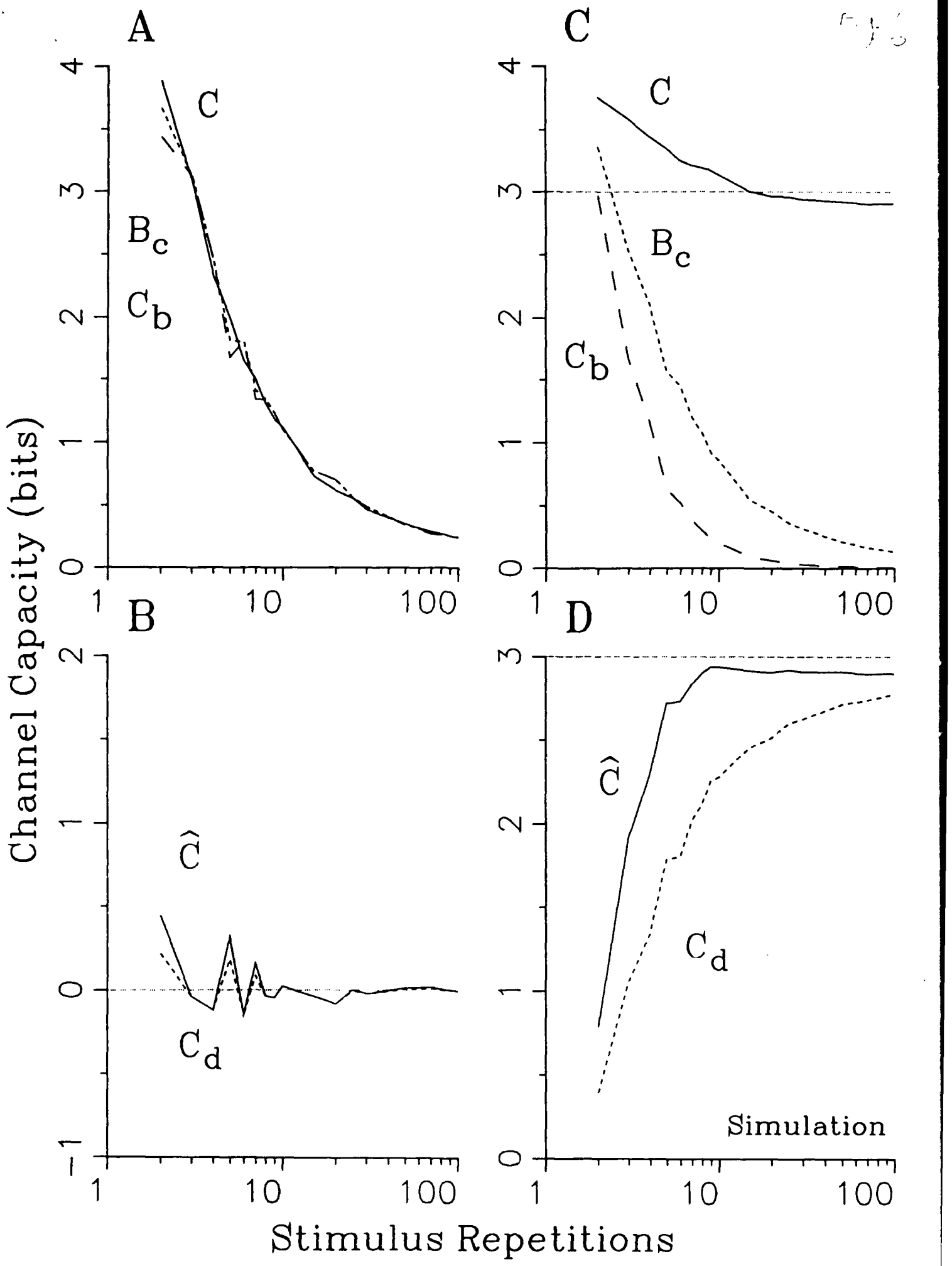

mg07a
T (bits)
0
1
2
3
4
5
10
100
1000
10000
Points Per Cloud
5-D
4-D
3-D
2-D
1-D

Fig 3


% Change in $\hat{T}$
5
4
3
2
1
0
−1
−2
−3
−4
−5
4−D
2−D
1−D
3−D
mg06
0
5
10
15
20
25
30
Bootstrap Repetitions


A
C
B
D
T
B
T_b
$\hat{T}$
T_d
Simulation
Transmitted Information (bits)
Stimulus Repetitions
0
1
2
3
4
−1
1
10
100


A
T
$\gamma = 1$
$\gamma = 3$
T_b
B
$\hat{T}$
$\gamma = 3$
$\gamma = 1$
Simulation
Transmitted Information (bits)
Stimulus Repetitions
0
1
2
3
4
1
10
100


A
C
B_c
C_b
C
B_c
C_b
B
D
Ĉ
C_d
Ĉ
C_d
Simulation
Channel Capacity (bits)
Stimulus Repetitions
1
10
100
0
1
2
3
4
−1

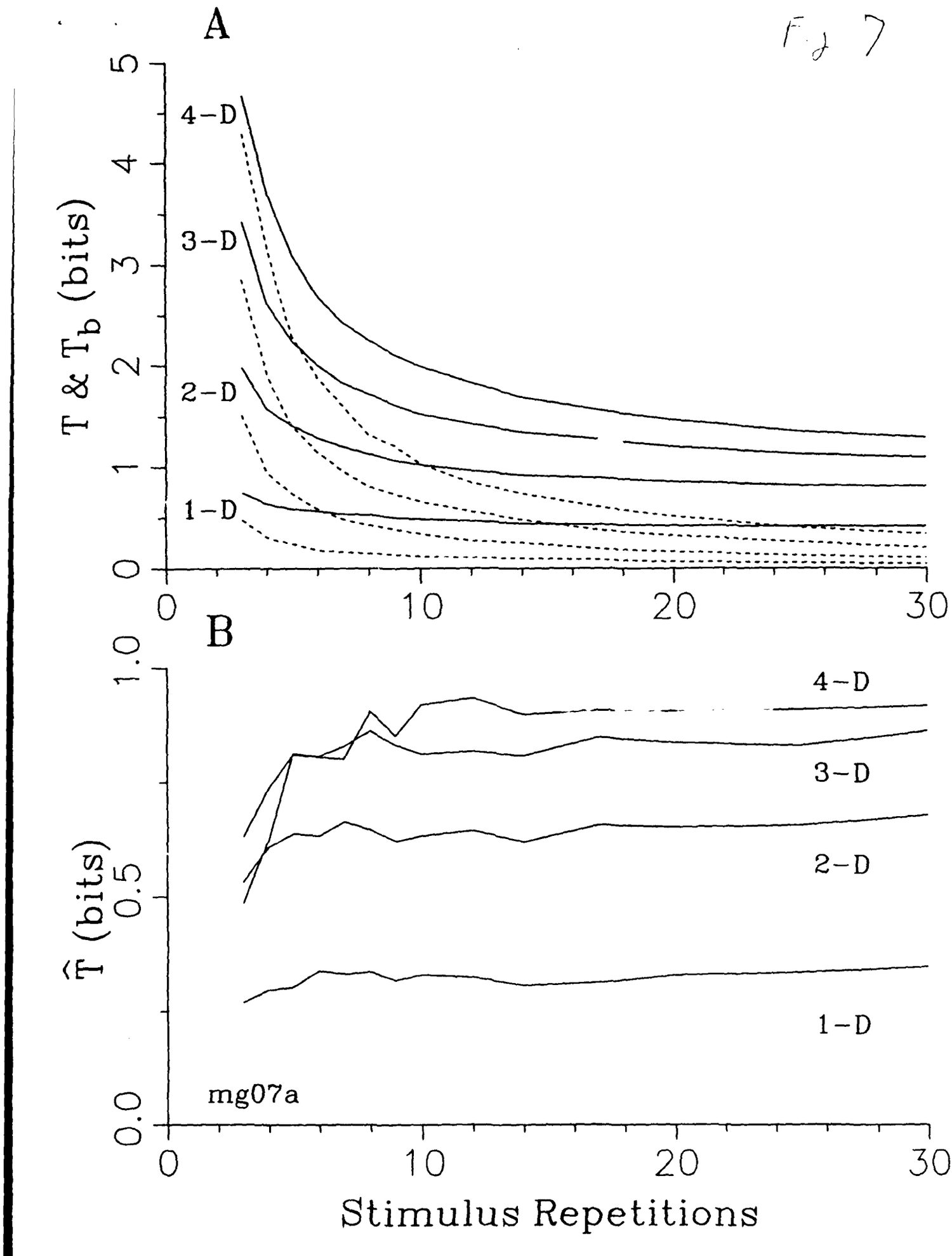

A
Fig 7
T & T_b (bits)
4-D
3-D
2-D
1-D
0
1
2
3
4
5
0
10
20
30
B
$\hat{T}$ (bits)
4-D
3-D
2-D
1-D
mg07a
0.0
0.5
1.0
0
10
20
30
Stimulus Repetitions

Fig 8

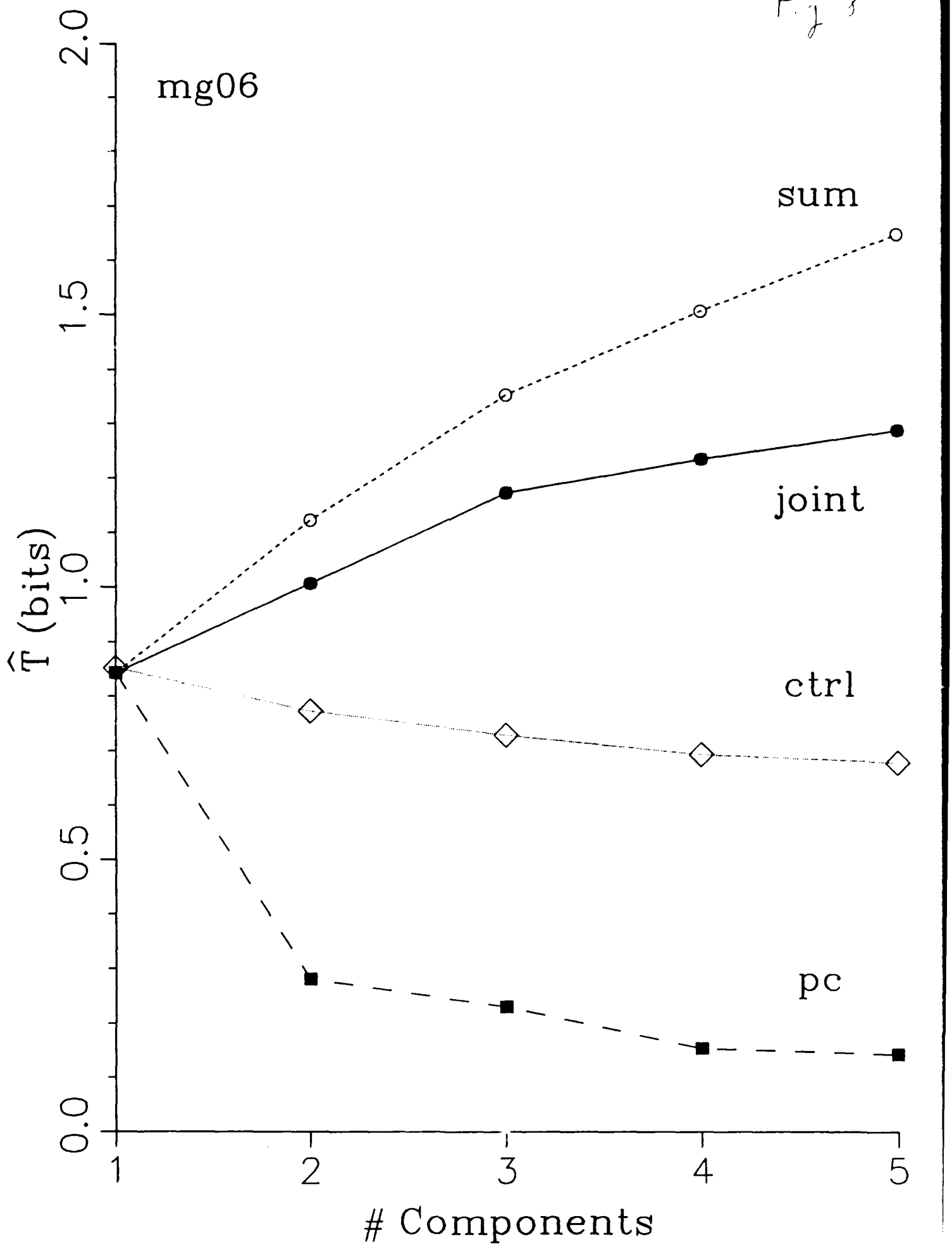
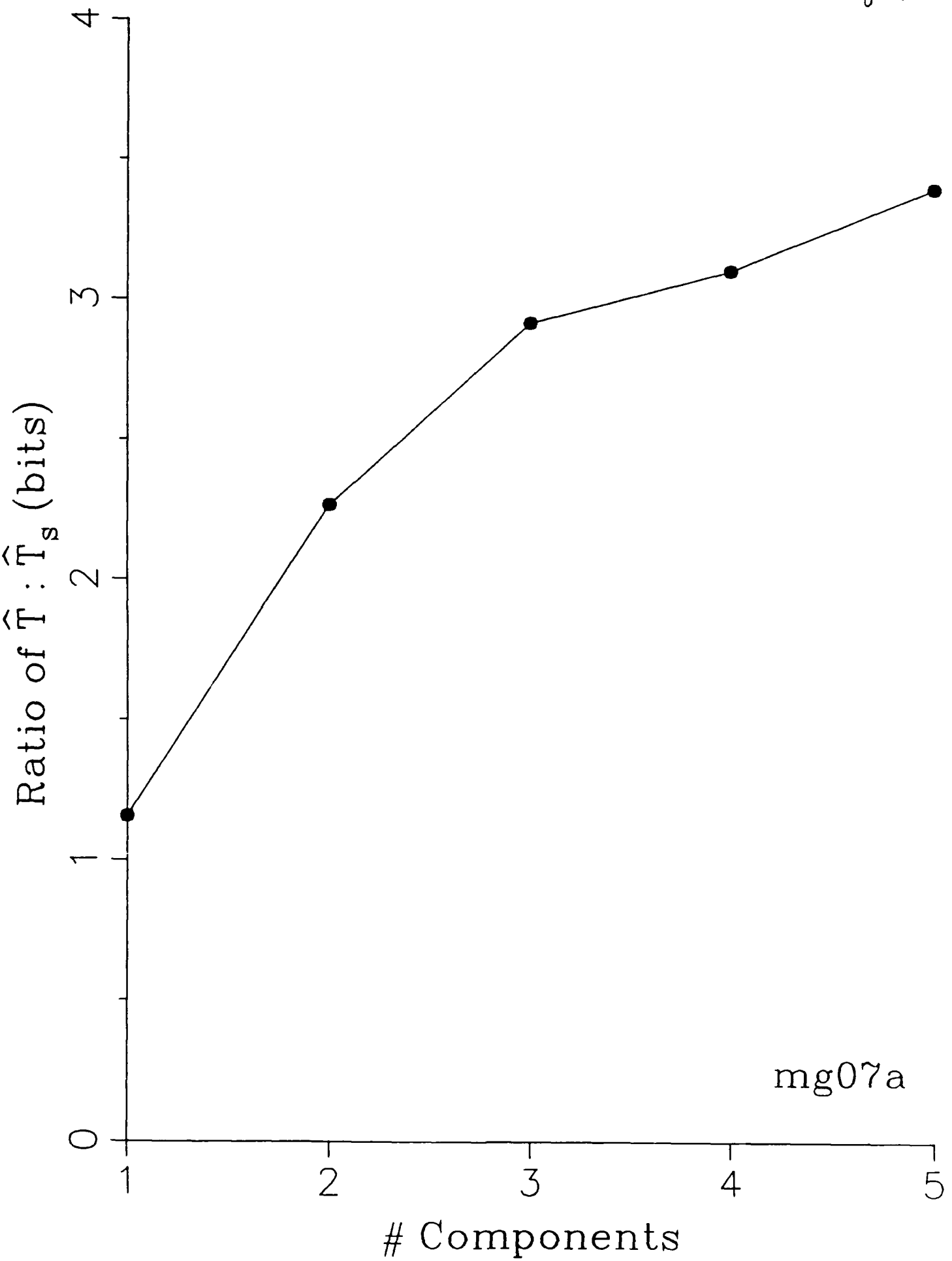

mg06
sum
joint
ctrl
pc
$\hat{T}$ (bits)
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
Components

Fig 9



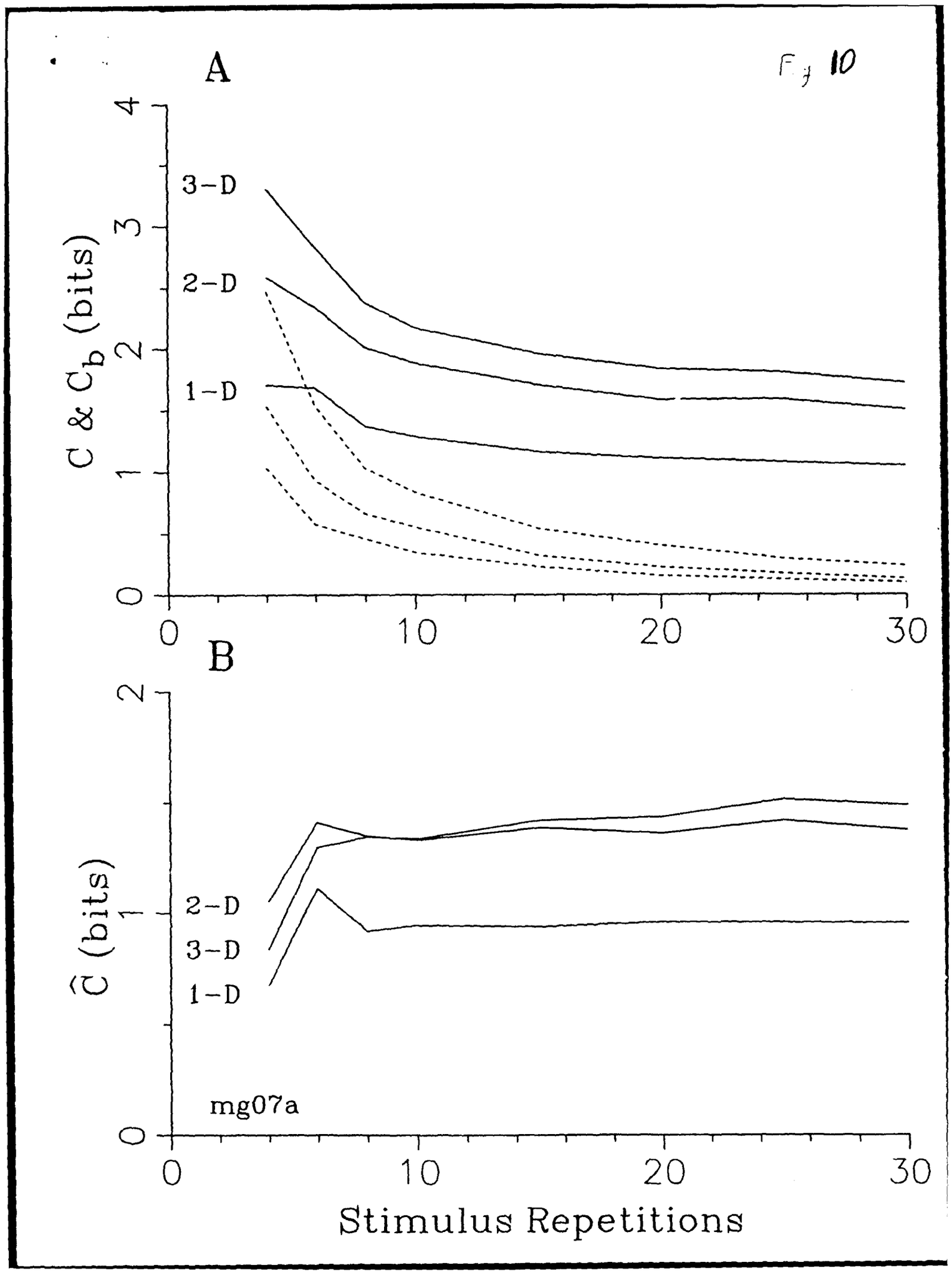

A
Fig 10
C & C_b (bits)
3-D
2-D
1-D
0
1
2
3
4
0
10
20
30
B
$\hat{C}$ (bits)
2-D
3-D
1-D
mg07a
0
1
2
0
10
20
30
Stimulus Repetitions

Fig. 11